Appetitive augmental functions and common physical properties in a pain-tolerance metaphor: An extended replication

Rosina Pendrous, Dr Lee Hulbert-Williams, Dr Kevin D. Hochard, and Professor Nicholas J. Hulbert-Williams

Author affiliation: Centre for Contextual Behavioural Science, School of Psychology, University of Chester, UK

[Accepted for publication in the Journal of Contextual Behavioural Science]

https://doi.org/10.1016/j.jcbs.2020.02.003

Correspondence concerning this article should be addressed to Rosina Pendrous, ropendrous@gmail.com; School of Psychology, University of Chester, Parkgate Road, Chester, CH1 4BJ.

Declarations of interest: none

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Word count: 6174

Highlights

- Additional contextual cues did not alter a metaphor’s effect on pain tolerance
- Original study’s findings were not replicated
- Highlights importance of testing metaphor use in ecologically valid settings

Abstract

Relational frame theory claims that the tacit understanding of metaphorical language rests upon our ability to derive relations based on relevant contextual cues; with metaphor aptness being a function of learning history and the number and nature of contextual cues presented. Recent experimental research has explored whether metaphor aptness plays a role in changing behaviour. Sierra et al. (2016) demonstrated that the presence of common physical properties (herein *common properties*; "cold") within a perseverance metaphor increased pain tolerance to the cold pressor task. When the metaphor also specified appetitive augmental functions (herein *augmentals*; "something important to you"), pain tolerance also increased. We tested the replicability of these findings under more stringent conditions, using a stratified (by sex) double-blind randomised-controlled experimental design. Eighty-nine participants completed baseline measures of psychological flexibility, cognitive fusion, generalised pliance, and analogical reasoning ability. Participants were then allocated to a pre-recorded audio-delivered metaphor exercise containing either: (i) common properties; (ii) augmentals; (iii) both; or (iv) neither (control condition). Participants completed the cold pressor task before and after intervention. We found no change in pain tolerance following intervention in any condition. Given potential implications for apt metaphor use for changing behaviour, further work is required to establish why the original study's findings were not replicated, to identify boundary conditions for the putative effect, and test metaphor use in ecologically valid settings.

Key words: *metaphors; analogy; relational frame theory; replication*

Introduction

Metaphorical language is ubiquitous, though somewhat difficult to define. Common definitions refer to the non-literal nature of metaphorical language (e.g., Oxford English Dictionary, 2018), though this merely shifts the epistemic burden to another term. Entire theses have been written in an attempt to understand the means by which metaphorical language manages to convey the speaker's meaning despite the words being used in an unconventional manner (e.g., Camp, 2003). Despite these difficulties, a number of assertions can be made with relative confidence. Metaphors are a method of symbolic and tacit representation that may otherwise be difficult to convey by literal language. Although metaphorical usages are ubiquitous to human language (e.g., Boroditsky, 2001), they become particularly useful when describing psychological phenomena. In the therapeutic setting, it has been suggested that metaphors help to convey and formulate ideas as to the source of a client's issues in cognitive behavioural therapies (Stott, Mansell, Salkovskis, Lavender, & Cartwright-Hatton, 2010), particularly to challenge maladaptive thinking styles and address biases or behaviour (Blenkiron, 2005). In "third wave" behavioural therapies, such as Acceptance and Commitment Therapy (ACT; Hayes, Strosahl, & Wilson, 1999), metaphors are used not least to deliteralise psychological content that might perhaps be problematic (Foody et al., 2014). Given their clinical utility and widespread use, it is important that we develop a clearer idea of how metaphors are understood by the listener, and especially, how and when metaphors used properly in the intervention setting can lead to behaviour change.

One's ability to understand metaphors is usually tested using analogical reasoning tests. An analogy, of course, can be explicit and conventional, and therefore non-metaphorical in manner. The four-term equality of proportion, "A is to B as C is to D" (often written symbolically as A:B::C:D) has been considered a typical structure for measuring analogical reasoning ability (Goswami, 2001) for many years. Questions like “dog (A) is to animal (B) as apple (C) is to D”, have been used as a metric for intelligence quotients tests, given the suggestion that analogical reasoning is an indicator of higher order cognitive ability (Stewart, 2016). Many metaphors have an implicit analogical structure, as we will later see.

Cognitive theorists have modelled analogical reasoning in many ways, usually dichotomising these approaches into those of ‘comparison’ (e.g., Gentner, 1983) and those of ‘categorisation’ (e.g., Glucksberg & Keysar, 1990). The overall view from cognitive science is that analogical comparison involves the ‘mapping’ of learned information that is familiar to us from one existing domain in memory (the “vehicle”) to another domain to be explained (the "target"; Gentner, 1983; Stewart & Barnes-Holmes, 2001). Different views exist therefore regarding how humans derive an analogical comparison and gain insight into the target as a result. Is this ability of the listener’s affected by features of the analogy, and if so, which features? To date, cognitive theorists have pointed to two features: aptness and conventionality (Jones & Estes, 2006).

From a Relational Frame Theory (RFT) perspective, metaphor use involves the abstraction of the functional properties of two arbitrary relational networks (the “vehicle” and the “target”), coordinated by contextual cues that imply sameness. For example, if the speaker says that “anger is a volcano”, the form of the sentence, including the words “is a” can act as contextual cues that indicate coordination between the target “angry” and the vehicle “volcano”. By discriminating the formal relation via coordination of the stimuli, a transformation of stimulus function occurs, whereby angry and volcano become functionally equivalent (e.g. unpredictable, explosive) within certain bounds dictated by the context. RFT thus attempts to analyse analogies (and thus many usages of metaphorical language) as relations amongst relations. This analysis has received some support, particularly from experimental work teaching such metaphors using the means implied by RFT (Barnes, Hegarty, & Smeets, 1997; Ruiz & Luciano, 2015; Stewart & Barnes-Holmes, 2001).

RFT research has begun to explore judgements of analogy aptness. In one computerised study (Ruiz & Luciano, 2015), 20 participants learned to respond to the analogical tasks in a match-to-sample format, and then were trained on two separate relational networks consisting of three equivalence classes. When the analogies contained common physical properties (nodes with colour spots), the analogy was judged as more apt (judged by participants selecting this option) than when there were no common physical properties. This suggests that analogies are more apt when there are common formal (physical) properties between relations (see Hayes, Barnes-Holmes, & Roche, 2001; Stewart & Barnes-Holmes, 2001).

In Acceptance and Commitment Therapy, as in many other approaches to psychological intervention, metaphors are used to explain concepts to clients and in an attempt to bring about behaviour change (see Hayes, Strosahl, & Wilson, 1999; Foody et al., 2014). Recent laboratory analogue research has begun to explore how metaphors might best be constructed so as to bring about behaviour change in the most reliable manner. In one study (Sierra et al. 2016), 80 undergraduate students completed one of four conditions (a 2×2 factorial design) using the Swamp metaphor (Hayes et al., 1999). This metaphor is intended to teach psychological flexibility, which is defined by Sierra et al. (2016, p. 267) as "the generalized repertoire of framing ongoing behavior in hierarchy with the deictic I (i.e., observing and taking distance from the ongoing behavior), which typically reduces the discriminative functions of ongoing behaviour and allows the derivation of rules that specify appetitive augmental functions (i.e., valued directions) and the behavior that is in accordance with them." Sierra and colleagues used a version of the metaphor which describes persevering through short-term unpleasantness (either a "cold" or a "filthy" swamp, depending on presence or absence of common properties) to get to something on the other side of the swamp in the long-term, specified as important or not (specification of augmentals or not). First, participants completed three self-report baseline measures: (i) psychological flexibility, (ii) cognitive fusion, (iii) generalised pliance, and then completed the cold pressor task for the first time to gain scores of the two main dependent variables: (i) baseline pain tolerance (time in the cold pressor), and (ii) pain intensity (measured on a visual analogue scale). In line with the RFT analysis of metaphors, the authors hypothesised that the inclusion of common properties (descriptions of a "very cold" swamp, given that the outcome measure involved cold tolerance), and the inclusion of appetitive augmental functions (described as "augmentals" hereinafter, like "something very important to you" is on the other side of the "cold" water), would increase tolerance in a subsequent cold pressor task. The findings indicated that when the metaphor included either common properties or augmentals, pain tolerance was increased. Self-reported pain intensity did not change regardless of condition. This study provided preliminary evidence to suggest that the use of an RFT analysis of metaphor aptness may useful in influencing pain tolerance.

Recently, Sierra et al.'s work has been partially replicated and extended by Criollo, Díaz-Muelle, Ruiz, and García-Martín (2018). The same design and procedures were used, however, each condition had three functionally equivalent metaphors (allowing for better abstraction of the behavioural rule) with each of the components specified above (i.e., Condition A had three metaphors with both augmental functions and common properties; Condition B had three metaphors with only common properties). Using Bayesian inferences, it was found that presence of common properties were the only factor which had an influence in increasing pain tolerance, whereas having multiple examples showed no effect (Criollo et al., 2018).

The current study

Research on the nature of metaphor choice and content is especially important for ACT: in the clinical setting, therapists will likely individualise metaphors to the particular client/behaviour, and in research settings, and in group-based delivery, we often manualise intervention content (including metaphor choice) which may limit the degree to which metaphors can be chosen to coincide with a participant's specific learning history.

The current study is an extended direct replication of Sierra et al. (2016). An *a priori* plan of all stages of the experiment, including the analysis plan, was preregistered (see https://osf.io/p2hwv/). We made the following changes from the original study to test the reproducibility of these findings under more stringent conditions. First, the computerised software PsychoPy (Peirce et al., 2019) was used to automate the task. This was in order to reduce potential experimenter effects and facilitate a truly double-blind design. Second, given sex differences often reported in studies using the cold-pressor task (Mitchell, MacDonald, & Brodie, 2004), participants were randomised by condition, but the ratio of males and females balanced across all conditions. Third, the original metaphor scripts from Sierra et al. (2016) were edited in the following ways:

(i) all conditions were made equal in both word count, audio length, and number of qualifying words (e.g., "very", "awful") and adjectives (e.g., "cold", filthy"), and

(ii) some phrases were altered to make the scripts flow more naturally, following a translation from Spanish to English. These were so the conditions were standardised and so the conditions which include no common properties did not include an analogy as descriptions of the swamp ("smells like a sewer") within the metaphor as it is unclear what effect this additional analogical relation would have.

Fourth, a measure of analogical reasoning ability, the Wechsler Adult Intelligence Scale®—Third Edition (WAIS-III; Wechsler, 1997) Similarities Subtest of the Verbal Subscale, was used at baseline. It was expected that one's ability to understand analogy, and derive similarity within the metaphor, may influence the strength of the metaphor's effect on pain tolerance, as recommended in the original paper (Sierra et al., 2016). It was therefore important to statistically control for how well one can understand analogies.

We hypothesised that there will be an increase in pain tolerance when common physical properties are included, and when appetitive augmental functions are specified. Finally, that there will be an additive effect of pain tolerance of both common physical properties and appetitive augmental functions.

Method

Participants

A priori sample size was determined using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) based on effect sizes of η^2 = .091 as reported in Sierra et al. (2016), and found that 88 participants would provide the appropriate level of power (1-β = .80) to detect main effects with a mixed ANOVA, using a of threshold of α = .05. Eighty-nine staff and students (77.5% female) aged between 18-59 (*M* = 24.69, *SD* = 8.85) were recruited from a UK university, including via the student participant pool. Incentives (for students only) were participant pool credits and, to facilitate recruitment over the quiet summer period, five pounds (about 6.5 US dollars). Recruitment methods included posters, emails, and advertisements on the participant pool system. Some exclusion criteria were applied, following health and safety precautions associated with the Cold Pressor Task: a history of cardiovascular disorder or Reynaud's disease, seizures or fainting, frostbite, and so on. Participants were also excluded if they exceeded 300 seconds (five minutes) on the cold pressor before intervention.

This study was approved by the [IRB name removed for anonymous review] School of Psychology ethics committee.

Design and Variables

A preregistered double-blind stratified (by sex) randomised controlled experimental mixed design was implemented. As in the original (Sierra et al. 2016) study, the three independent variables were:

(i) presence of common physical properties (yes/no; known hereinafter as 'common properties'),

(ii) presence of appetitive augmental functions (yes/no; 'augmentals'), and

(iii) timepoint (pre/post).

The two dependent variables were:

(i) pain tolerance time (in seconds) to the cold pressor was measured using a stopwatch, and

(ii) perceived pain strength (akin to "pain intensity"), measured using an adapted version of the Borg CR-10 (1985) scale, ranging from nothing at all (0) to absolute maximum (10). Meta-analytic research has found this to be a valid measure (Chen, Fan, & Moe, 2002), and has been found to be slightly more efficient than visual analogue scales (Neely et al., 1992), as in Sierra et al. (2016). We chose this test as it allows the participants to use any number (e.g., 0.3, 6.7), including 11 (i.e. the worst pain experienced yet, as possible with the cold pressor task). This is standard to the test, but mean scores were calculated.

Measures and Materials

Participants were tested individually in a small laboratory. The open-source graphical user interface on PsychoPy v.1.84.2 (Peirce et al., 2019) was used to build the experiment on the computer, presenting the psychometrics and delivering the pre-recorded audio scripts via headphones. The cold pressor machine (Jeio Tech Refrigerating Bath Circulator RW-2025G) provides a 15x20 inch basin filled water and circulated at a constant temperature of 3°C.

In addition to basic demographic questions (age and sex), participants completed the following four psychometric scales at baseline. The first is from a well-known intelligence test. The remainder are self-report questionnaires with 7-point Likert-type response options, ranging from never true (1) to always true (7), with higher scores indicative of more of the relevant trait.

Analogical reasoning ability. Analogical reasoning was measured using the similarities subtest of the verbal subscale to the WAIS–III (Wechsler, 1997). Participants were asked to state how two objects or concepts are similar (for example, “how are ‘fork’ and ‘spoon’ alike?"), and then scored (0, 1, 2) in accordance with predefined criteria. The typical structure of these questions arguably encourages participants to derive relations between relations (analogical/metaphorical) of the stimuli presented. This provided an overall score (max 33) of how well one can understand similarities in words and concepts across 19-items. A higher score indicated higher ability to understand similarity within two concepts. It makes no sense to provide Cronbach’s alpha for ability tests as the response options are dichotomous and items are varied and scaled with respect to difficulty, however, this sub-scale is taken from an extremely well-validated test.

Psychological flexibility. Psychological flexibility was measured across seven items using the *Acceptance and Action Questionnaire- II* (AAQ-II; Bond et al., 2011). Reverse scoring was applied to obtain a measure of flexibility. The measure was selected to assess the extent to which individuals are willing to experience difficult internal events while continuing to behave in alignment with values; an example item: “I am afraid of my feelings”. Cronbach’s alpha for this scale in in the original study was α = .87. In the present study this was AAQ-II was α = .89.

Cognitive fusion. Cognitive fusion was assessed across seven items using the *Cognitive Fusion Questionnaire* (CFQ; Gillanders et al., 2014). This measure was selected to assess the attachment between language and behaviour; an example item: “I struggle with my thoughts”. Cronbach’s alpha for this scale in the original study was α =.90. In the present study this was α = .93.

Generalised pliance. Generalised pliance was assessed across 18 items using the *Generalised Pliance Questionnaire* (GFQ; Ruiz, Suárez-Falcón, Barbero-Rubio, & Flórez, 2018). This measure was selected to assess how much of the participants’ typical behaviour is mediated by social approval, not least due to evident demand characteristics influencing scores on the cold pressor task in the presence of an experimenter (see Roche, Forsyth, & Maher, 2007); an example item: “In order to be happy, I need people to value me”. Cronbach’s alpha for this scale in the original study was α = 0.92. In the present study this was α = .95.

Task-specific questions. Participants were asked four additional questions: (i) “How likely would you be to use this technique in real life?”, to assess subjective external validity (the technique was not re-described); (ii) “Is English your first language?”, given cultural differences in understanding metaphorical language (e.g., Boroditsky, 2001); (iii) “Have you ever used the cold pressor before?”, to assess possible practice effects; and (iv) “How knowledgeable are you about ACT or RFT?”, to assess potential knowledge of the underlying aims of the study (using abbreviations). Responses to (ii), and (iii) were binary (yes/no); responses to (i) and (iv) were on a three-point Likert-scale ranging from “not at all” to “a lot”.

Procedure (see Figure 1)

Pre test. Upon arriving at the laboratory, all participants were asked to read an information sheet and provide written consent in paper format. Before completing the analogical task with the participant, the experimenter provided verbal instructions. Participants were then asked to turn to the computer and complete the self-report questionnaires and basic demographics on-screen. Participants were given clear instructions to standardise the task (e.g., to submerge their hand up to a specific point on the wrist). Participants recorded pain strength on the computer.

Experimental manipulation. Participants were randomised using an Excel random number generator. Double-blind randomisation was achieved by the conditions being renamed by the third author who then had no involvement with the data collection. Sex was balanced across all conditions through a re-randomisation process should the allocated group have two more males than any other group. This process was fully automated in advance. Participants then listened to a scripted speech describing the Swamp metaphor, differing only in whether the condition specified common properties and/or augmentals (see below for the scripts; the bold and italicised text indicate what differed between conditions).

> You have just completed the cold-pressor task. Do you remember the sensation you felt in your hand while you were doing the task (pause of 15s)? As you know, the aim of this experiment is to analyse which strategies people with chronic pain could use to obtain the things important for them even though they are experiencing pain. Your participation in this experiment is important because it could contribute to the quality of life of individuals living with chronic pain. We are not expecting any results

in particular, anything you do is OK for us. We only ask that you do the task honestly and try to follow the next exercise. As soon as you're ready, close your eyes, and really try to imagine you are at the edge of a big swamp. The other side of the swamp is very far away, and it would take you several minutes to get there. On the other side of the swamp, ***there is the most important thing to you - something that really excites you, or something that you often think about [in Conditions B and D, this is "there is a landscape that is exactly the same as the one you are seeing on your side"].*** Please, let yourself think for a few seconds what would be on the other side of the swamp and the emotion that would drive you to get there (pause of 30s). Now, we are going to ask you a brief question. If you could open your eyes for just a moment (pause of 5s). The water of the swamp is ***very cold [in Conditions C and D the water was described as "disgusting– it smells awful"]***, and when you look to the other side, you realise that the only way to get there is to cross the swamp by swimming. It would take you five minutes to get to the other side. The farther you swim across the swamp, the more ***cold [in Conditions C and D, this is "disgust*"]** you would feel, but you know that you would be much closer to ***this thing that is so important for you [in Condition B and D, this is "the other side"].*** You would also know that ***cold [in Conditions C and D this is "disgust"]*** is something you would feel momentarily, something uncomfortable that it makes sense to feel for a few minutes ***because on the other side is the most important thing for you [in Conditions B and D this is "to reach the other side"].*** Please, let yourself imagine the feeling that you would have swimming in the swamp while going to the other side and the feeling that you would have seeing the other side closer (pause for 15s). Would you stand at the edge of the swamp looking at ***how the most important thing*** for you fades away on the other side ***[in Conditions B and D this is "landscape from the other side]*** or would you jump into the water and swim despite the discomfort of the ***cold***? ***[in Conditions C and D this is "disgust"]*** (pause of 10s). This part of the experiment has now finished, and we are going to ask you to do the cold pressor task again. Please try to put into practice what you have just imagined. Please call the experimenter to indicate that you're ready to try the cold pressor again.

Post test. All participants were asked to complete the cold pressor task, and asked to rate the pain strength, for the second and final time. Upon completion, participants were provided with written and verbal debriefing.

INSERT FIGURE 1 ABOUT HERE

Deviations from preregistration

In the preregistration, we hypothesised a decrease in task aversiveness (asked with a question "would you be willing to retake the cold pressor task?" with a yes/no response), an additional dependent variable. Initially, we had planned to assess this pre and post, however, we removed it from the pre-test assessment as we realised that an answer of 'yes' would in fact simply be another way for participants to opt-out from the study. For information on the results using this dependent variable, please see the output on our OSF page.

Results

All data preparation and analyses were carried out using SPSS v.24. Five participants were excluded due to running up against the 300 second (five minute) ethical constraint on the cold pressor, leaving an effect N of 84. Missing data were excluded on a listwise basis.

Confirmatory analyses

In line with our preregistered analysis plan (see https://osf.io/p2hwv/), four one-way ANOVAs were run to check for baseline differences between condition on each trait measure. There were no significant baseline differences between conditions on: psychological inflexibility, $F(3, 79) = .120$, $p = .948$; cognitive fusion, $F(3, 79) = .292$, $p = .831$; and, generalised pliance, $F(3, 79) = .105$, $p = .957$. The difference between conditions at baseline with respect to analogical reasoning ability came close to usual statistical significance cut-offs, $F(3, 80) = 2.702$, $p = .051$. Analogical reasoning was therefore included as a covariate in the following analyses given its theoretical relevance (see Table 1 for means and standard deviations [SDs]).

INSERT TABLE 1 ABOUT HERE

A 2×2×2 mixed ANCOVA tested the hypotheses that there will be an increase in pain tolerance (in seconds) when common properties only are included, when augmental functions are included, and when both are included. Analogical reasoning was used as a covariate given its theoretical importance and near-significant heterogeneity at the point of randomisation. With respect to pain tolerance, there was no significant main effect of time, $F(1, 79) = 1.509$, $p = .223$, $\eta_p^2 = .019$. There was a significant interaction between Time by Analogical Reasoning, $F(1, 79) = 4.663$, $p = .034$, $\eta_p^2 = .056$, but no significant interactions between Time and Common Properties, $F(1, 79) = .682$, $p = .411$, $\eta_p^2 = .009$; Time and Augmental Functions, $F(1, 79) = .239$, $p = .626$, $\eta_p^2 = .003$; Time and Common Properties and Augmental Functions, $F(1, 79) = 1.038$, $p = .311$, $\eta_p^2 = .013$. No condition displayed significantly greater levels of pain tolerance (see Table 2). It is worth noting that when analogical reasoning was not controlled for, there was only one significant effect — a main effect of time, $F(1, 80) = 24.197$, $p < .01$, $\eta_p^2 = .232$. The pattern of results was otherwise unchanged by the removal of the covariate.

INSERT TABLE 2 ABOUT HERE

The following analyses were not included in the preregistered protocol and are therefore strictly exploratory. A 2×2×2 mixed ANCOVA was run to explore whether there would be a change in pain strength (a self-reported rating) across conditions. Metaphors such as the one used in the current work are intended to increase psychological flexibility — the willingness to withstand difficult experiences for valued ends — and not to alter the subjective experience itself. Thus, we did not expect a change in self-reported pain strength. Analogical reasoning was used as a covariate given near significant differences at randomisation. There was no significant main effect of time on pain strength from Time 1 to

Time 2, $F(1, 57) = .484$, $p = .490$, $\eta_p^2 = .008$. There were no significant interactions between Time and Analogical Reasoning, $F(1, 57) = .008$, $p = .929$, $\eta_p^2 = .000$; Time and Common Properties, $F(1, 57) = .444$, $p = .508$, $\eta_p^2 = .008$; Time and Augmental Functions, $F(1, 57) = 3.101$, $p = .084$, $\eta_p^2 = .052$; Time and Common Properties and Augmental Functions, $F(1, 59) = .002$, $p = .968$, $\eta_p^2 = .000$. No condition displayed significantly greater levels of pain strength (see Table 2).

Confirmatory moderation analyses of the trait measures taken at baseline on Condition and cold pressor time difference (post time – pre time) were carried out in SPSS PROCESS v3.2 (Hayes, 2017) using indicator coding and bootstrapping (5000; $N = 84$). Condition D (control) was the reference category. No measure was a significant moderator on cold pressor time; psychological inflexibility, $R^2 = .081$, $F(7, 75) = .944$, $p = .478$; cognitive fusion, $R^2 = .065$, $F(7, 75) = .742$, $p = .638$ (see Table 3); generalised pliance, $R^2 = .045$, $F(7, 75) = .509$, $p = .825$; analogical reasoning: $R^2 = .130$, $F(7, 76) = 1.628$, $p = .140$ (see Table 4).

*** INSERT TABLE 3 ABOUT HERE ***

*** INSERT TABLE 4 ABOUT HERE ***

We also explored the other questions. Regarding whether participants would be likely to use this technique in real life, 8 participants (9.5%) responded “not at all”, 44 participants (52.4%) responded “somewhat”, and 31 participants (36.9%) responded “a lot”. Regarding previous knowledge of RFT/ACT, 50 participants (59.5%) responded “not at all”, 22 participants (26.2%) responded “somewhat”, and 11 (13.1%) responded “a lot”. These responses, however, may have been skewed by participants knowledge of what these commonly used abbreviations stand for. English was 66 participants' first language (78.6%), and 64 participants (76.2%) had not done the cold pressor before.

Discussion

The present study was a preregistered extended replication of Sierra et al. (2016) to assess the reproducibility of the findings relating to metaphor aptness. Sierra and colleagues used a 2×2×2 design to test whether the inclusion of common properties in a metaphor describing persevering through a "very cold" swamp lead to increased pain tolerance to a very cold behavioural task, and whether specifying appetitive augmental functions ("something very important to you") also increased pain tolerance. Our extended direct replication using largely the same procedures but with slightly improved materials, was carried out under more stringent conditions. These included automating the intervention delivery so as to achieve double blinding (to rule out explanations relating to demand characteristics and the placebo effect), stratifying the randomisation process by sex (to control for the known sex differences shown in pain research), and balancing script lengths (so as to rule out explanations to do with dose-related effects).

The key findings in Sierra et al. (2016) were not directly replicated. Pain tolerance did not significantly increase following administration of the metaphor in any condition. This lack of effect was observed while statistically controlling for a proxy measure of analogical reasoning ability. Whether the metaphor specified augmental functions or common properties, both (an additive effect), or neither, had a negligible effect on pain tolerance. According to RFT, understanding metaphorical language depends on our ability to derive arbitrary relations based on relevant contextual cues. Common properties ("cold") in metaphors between the cold pressor task (also "cold") were said to act as additional contextual cues that facilitate analogy derivation, which could reduce cognitive burden associated with responding to metaphorical language. It was also expected that the inclusion of augmentals (verbal rules that describe the consequences of a behaviour) may also increase pain tolerance as this transforms the function of the behaviour. By specifying “something important to you”, this may give the participants a reason (a value) to tolerate the cold water; this was not the case in the present study.

We did not hypothesise that there would be a change in pain strength following either condition, since this was non-significant in the original study when using a visual analogue scale (Sierra et al., 2016). Similarly, exploratory analyses found no changes in pain strength across conditions, when using a validated measure of perceived pain strength (Borg, 1985). This is unsurprising as we would expect that acceptance-based interventions do not work to change the amount of pain experienced. Rather, and in the case of the current study, metaphors aim to teach acceptance and psychological flexibility – the ability to withstand short-term pain for the sake of longer-term values (Hayes et al., 2006). This finding is consistent with other research suggesting that there are no changes in pain intensity following mindfulness (e.g., Liu, Wang, Chang, Chen, & Si, 2013) or acceptance-based interventions (e.g., Kohl, Rief, & Glombiewski, 2012; Roche et al., 2007).

At a theoretical level, the current findings fail to support the theory, rooted in RFT, and originally tested in an analogue setting by Sierra et al. (2016), that the inclusion of common physical properties, and the specification of a proxy for a value (augmentals, “something very important to you”), would improve the ability of a metaphor to bring about behaviour change. While Sierra’s experimental design did take behavioural functions into account by asking participants to think about something important to them, it is possible that they might not have been as salient as would be the case in a therapeutic setting. Other computerised and experimental tasks involving the training and testing of analogical relations found supporting evidence for the assumption that analogical relations were judged more apt when there was greater similarity between the target and the vehicle, both from behavioural (Barnes et al., 1997; Ruiz & Luciano, 2015; Stewart & Barnes-Holmes, 2001) and cognitive (e.g., Jones & Estes, 2006) research, without assessing behaviour change. The original study (Sierra et al., 2016) was partially replicated more recently by Criollo et al. (2018), finding that the presence of common physical properties was associated with an increase in pain tolerance. However, given that the current study had many methodological strengths in comparison to these two studies (e.g. double-blind randomisation, stratification), suggesting that under more stringent experimental conditions, we do not yet have sufficient knowledge of the contextual factors affecting the power of metaphors to change behaviour.

Given that direct replications are essential to provide confidence in the reliability of findings (Open Science Framework, 2015) and the underlying idea of a hypothesis (Brandt et al., 2014; Simons, 2014), and given that few psychology journals explicitly state that they accept replications (Martin & Clarke, 2017), despite the ongoing replication crisis within our discipline, it is likely that many areas of psychological research are beset by the publication of false positives. As such, direct replications with high experimental rigour, are of vital importance also for CBS. Schäfer and Schwarz (2019) demonstrate that pre-registered studies typically report smaller effect sizes on average, which was also the case in our study.

At a practical level, metaphors are a tool used extensively in both “third-wave” therapies such as ACT (Hayes et al., 1999), and “second-wave” cognitive-behavioural therapies (Stott et al., 2010). In these settings, metaphors are said to aid understanding and provide an alternative perspective of psychological phenomena, such as cognitive biases (Blenkiron, 2005). The results of the current study tentatively suggest that in a well-controlled laboratory-based preparation, using an “analogue” setting described in Sierra et al. (2016), this specific metaphor does not appear reliably to promote pain tolerance to the cold pressor task, despite many participants (52.4%) suggesting that they would be somewhat likely to use this technique in real life. It is important to reemphasise that we often use generic metaphors in ACT research using manualised approaches, and therefore testing (and replicating) of specific metaphors is important.

The present work contributes negative findings regarding the effect of appetitive augmentals and common physical properties in a metaphor-based verbal intervention. Positive findings have been reported by Sierra et al. (2016). It may be that this failure to replicate is explained by the increased experimental controls in the current study. Of course, it may also be that the failure to replicate is an effect of one or more limitations in our own study design. It is therefore important to note four limitations and threats to validity. First, it could not be assessed whether the participants' hand had returned to normal temperature before the second assessment of pain tolerance (approx. five minutes), given that some researchers have suggested that the cold pressor may have a longer recovery time of approximately ten minutes (Edens & Gil, 1995). Second, the negative words in the conditions which did not include common properties ("filthy" instead of "cold") may have been aversive, a limitation also highlight Criollo et al.'s (2018) replication attempt. If "filthy" is considered more aversive than "cold" for the average participant, then the different conditions would not be testing whether differences in similarity to the task changes pain tolerance, but subjective aversiveness. Third, an *a priori* power calculation suggested that we needed 88 participants to see an effect; we recruited N = 89, but after removal of missing data in respect of the main dependent variables, the study was marginally underpowered (N = 84). A sample that leaves the study underpowered may also account for a possible inability to detect interaction effects with additional variables (Green, 1991), however the observed effect sizes are quite small. Finally, the sample had a large female bias (77.5%). While we were sure to control for the ratio of males within each condition, this limits generalisability to the general population outside of a lab-based setting.

To rectify the above limitations, future studies should firstly ensure that there is optimal recovery time (ten minutes or more) between measurements of the cold pressor. Secondly, to test whether differences in temperature across studies in part contributed to the failure to replicate, future direct replications could ensure similarity in temperature. Third, recruiting a larger sample would be useful to power analyses that specify more interactions and to better account for attrition and incomplete datasets. This may also increase the likelihood of having higher variability in key baseline measures, like analogical reasoning ability, to explore whether there is a minimum score for therapeutic techniques like metaphors to work. Finally, given that we did not assess those participants who were actively engaging with the metaphor task, and thus did not exclude those who were not, this could have in part caused the failure to replicate. Coaching and therapy would not be expected to work without active participation, which, in the case where intervention components are being tested, would mean that researchers should ensure that participants are motivated to engage legitimately. One way to rectify this could be to replicate the current study using no incentives (participant pool credits and/or money), so that we are testing data from those who are intrinsically 'motivated' to take part.

There are also several future directions to build upon previous research. First, state measures relevant to CBS have recently become available, such as the State Cognitive Fusion Questionnaire (Bolderston et al., 2018), as well as their trait variations. Using state measures pre and post could be useful to more sensitively assess changes in these constructs following the brief metaphors. Second, replicating the current study using a longitudinal design could be useful to test whether repeated use of the same metaphor leads to a change in pain tolerance. Third, future studies may wish to include more information that directly connects the metaphor to the task (e.g. by saying "During the task, try to imagine that the cold water is the swamp, and you are holding your hand in the water in order to get closer to what/who is really important to you"), or that is individualised for each participant as might be expected in a clinical setting. Finally, future studies testing the utility of metaphors for behaviour change could add a self-report measure to judge whether the metaphor is apt, as in other computerised work (Ruiz & Luciano, 2015).

Conclusion

Metaphorical language is often used in therapeutic settings to portray concepts or to provide new perspectives on psychological phenomena that is otherwise difficult to do using literal language. Research into the psychology of metaphors has also accumulated decades of empirical and theoretical evidence; the current preregistered study aimed to test the reproducibility of findings that suggest metaphors can be used to change behaviour using a behavioural model of language under a contextual-behavioural framework (Sierra et al. 2016). Under more stringent conditions, the key hypothesis was not supported as pain tolerance did not increase after any metaphor intervention. This suggests that further studies need to be carried out to establish the boundary conditions which obtain a potentially therapeutic metaphor effect. Longitudinal studies, replications carried out with greater methodological rigour, and broader tests of metaphor applicability in clinical samples and settings, are all necessary to explore theoretical reasons why these findings did not replicate. Within the context of our aims specifically, these findings further highlight the importance of pre-registered, transparent, replication studies within CBS.

References

Barnes, D., Hegarty, N., & Smeets, P. M. (1997). Relating equivalence relations to equivalence relations: A relational framing model of complex human functioning. *The Analysis of Verbal Behavior*, *14*(1), 57-83.

Blenkiron, P. (2005). Stories and Analogies in Cognitive Behaviour Therapy: A Clinical Review. *Behavioural and Cognitive Psychotherapy*, *33*(1), 45–59. https://doi.org/10.1017/S1352465804001766

Bolderston, H., Gillanders, D. T., Turner, G., Taylor, H. C., Ní Mhaoileoin, D., & Coleman, A. (2018). The initial validation of a state version of the Cognitive Fusion Questionnaire. *Journal of Contextual Behavioral Science*. https://doi.org/10.1016/J.JCBS.2018.04.002

Bond, F. W., Hayes, S. C., Baer, R. A., Carpenter, K. M., Guenole, N., Orcutt, H. K., … Zettle, R. D. (2011). Preliminary Psychometric Properties of the Acceptance and Action Questionnaire – II : A Revised Measure of Psychological Inflexibility and Experiential Avoidance. *Behavior Therapy*, *42*(4), 676–688. https://doi.org/10.1016/j.beth.2011.03.007

Borg, G. (1985). *An introduction to Borg's RPE-scale*. Mouvement Publications.

Boroditsky, L. (2001). Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time. *Cognitive Psychology*, *43*, 1–22. https://doi.org/10.1006/cogp.2001.0748

Chen, M. J., Fan, X., & Moe, S. T. (2002). Criterion-related validity of the Borg ratings of perceived exertion scale in healthy individuals: A meta-analysis. *Journal of Sports Sciences*, *20*(11), 873-899.

Criollo, A. B., Díaz-Muelle, S., Ruiz, F. J., & García-Martín, M. B. (2018). Common Physical Properties Improve Metaphor Effect Even in the Context of Multiple Examples. *The Psychological Record*, *68*(4), 513–523. https://doi.org/10.1007/s40732-018-0297-9

Edens, J. L., & Gil, K. M. (1995). Experimental induction of pain: Utility in the study of clinical pain. *Behavior Therapy*, *26*(2), 197–216. https://doi.org/10.1016/S0005-7894(05)80102-9

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *2*(39), 175–191.

Foody, M., Barnes-Holmes, Y., Barnes-Holmes, D., Törneke, N., Luciano, C., Stewart, I., & Mcenteggart, C. (2014). RFT for clinical use: The example of metaphor. *Journal of Contextual Behavioral Science*, 3(4), 305-313. https://doi.org/10.1016/j.jcbs.2014.08.001

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.

Gillanders, D. T., Bolderston, H., Bond, F. W., Dempster, M., Flaxman, P. E., Campbell, L., … Remington, B. (2014). The Development and Initial Validation of the Cognitive Fusion Questionnaire. *Behavior Therapy*, *45*(1), 83–101. https://doi.org/10.1016/j.beth.2013.09.001

Glucksberg, S., & Keysar, B. (1990). Understanding Metaphorical Comparisons: Beyond Similarity. *Psychological Review*, 97(1), 3.

Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, *26*(3), 499–510.

https://doi.org/10.1207/s15327906mbr2603_7

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory : a post-Skinnerian account of human language and cognition*. Springer Science & Business Media.

Hayes, S. C., Strosahl, K., & Wilson, K. G. (1999). *Acceptance and Commitment Therapy: An experiential approach to behavior change*. New York: The Guilford Press.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and Commitment Therapy: Model, processes and outcomes. *Behaviour Research and Therapy*, *44*, 1–25. https://doi.org/10.1016/j.brat.2005.06.006

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*(1), 18–32. https://doi.org/10.1016/J.JML.2006.02.004

Kohl, A., Rief, W., & Glombiewski, J. A. (2012). How effective are acceptance strategies? A meta-analytic review of experimental results. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(4), 988–1001. https://doi.org/10.1016/J.JBTEP.2012.03.004

Liu, X., Wang, S., Chang, S., Chen, W., & Si, M. (2013). Effect of Brief Mindfulness Intervention on Tolerance and Distress of Pain Induced by Cold-Pressor Task. *Stress and Health*, *29*(3), 199–204. https://doi.org/10.1002/smi.2446

Martin, G. N., & Clarke, R. M. (2017). Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices. *Frontiers in Psychology*, *8*, 523. https://doi.org/10.3389/fpsyg.2017.00523

Mitchell, L. A., MacDonald, R. A. ., & Brodie, E. E. (2004). Temperature and the cold pressor test. *The Journal of Pain*, *5*(4), 233–237. https://doi.org/10.1016/j.jpain.2004.03.004

Neely, G., Ljunggren, G., Sylven, C., & Borg, G. (1992). Comparison between the Visual Analogue Scale (VAS) and the Category Ratio Scale (CR-10) for the evaluation of leg exertion. *International Journal of Sports Medicine, 13*(02), 133.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 1–9. https://doi.org/10.3758/s13428-018-01193-y

Roche, B., Forsyth, J. P., & Maher, E. (2007). The Impact of Demand Characteristics on Brief Acceptance- and Control-Based Interventions for Pain Tolerance. *Cognitive and Behavioral Practice, 14*(4), 381–393. https://doi.org/10.1016/J.CBPRA.2006.10.010

Ruiz, F. J., & Luciano, C. (2011). Cross-domain analogies as relating derived relations among two separate relational networks. *Journal of the Experimental Analysis of Behavior*, *95*(3), 369-385. https://doi.org/10.1901/jeab.2011.95-369

Ruiz, F. J., Suárez-Falcón, J. C., Barbero-Rubio, A., & Flórez, C. L. (2018). Development and initial validation of the Generalized Pliance Questionnaire. *Journal of Contextual Behavioral Science*. https://doi.org/10.1016/J.JCBS.2018.03.003

Sierra, M. A., Ruiz, F. J., Flórez, C. L., Riaño Hernández, D., & Luciano, C. (2016). The role of common physical properties and augmental functions in metaphor effect. *International Journal of Psychology and Psychological Therapy*, *16*(3).

Stewart, I. (2016). The fruits of a functional approach for psychological science. *International Journal of Psychology*, *51*(1), 15–27. https://doi.org/10.1002/ijop.12184

Stewart, I., & Barnes-Holmes, D. (2001). *Understanding Metaphor: A Relational Frame Perspective*. *The Behavior Analyst*, *24*(2), 191.

Stott, R., Mansell, W., Salkovskis, P., Lavender, A., & Cartwright-Hatton, S. (2010). *Oxford guide to metaphors in CBT: Building cognitive bridges*. Oxford University Press.

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Tables and Figures

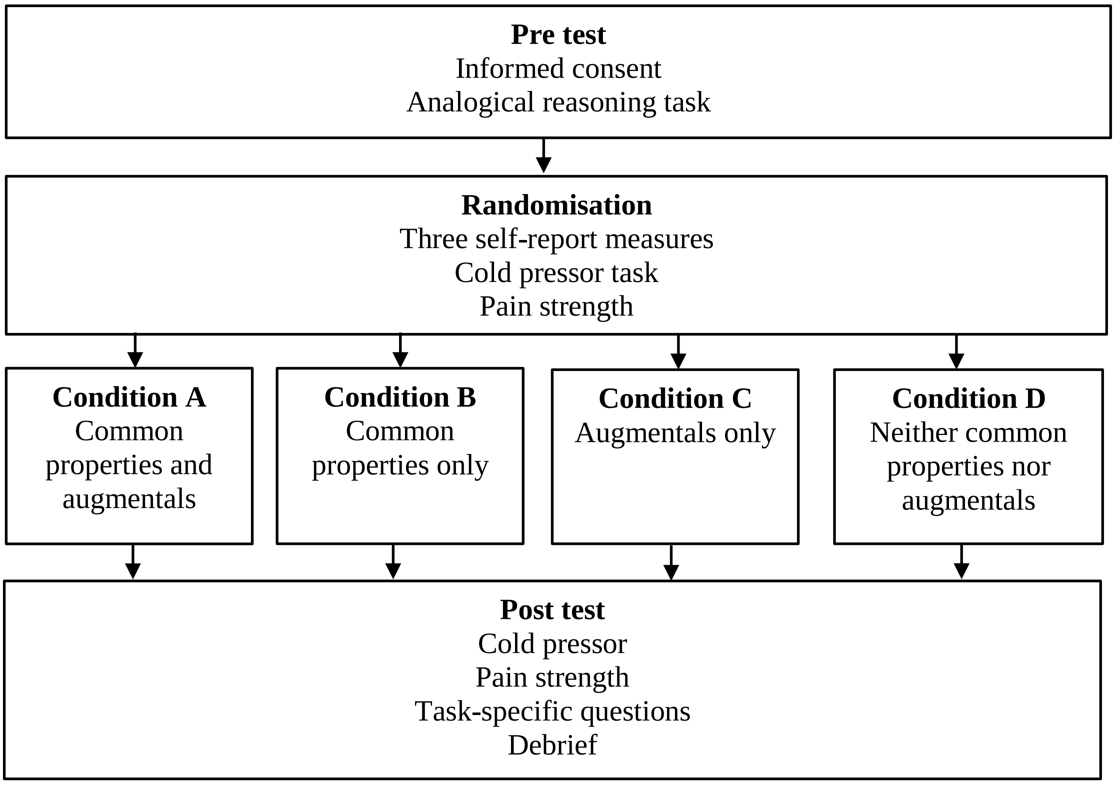

Figure 1. A visual representation of the experimental procedure.

Table 1. *Means, standard deviations/frequency (%), and N for study variables per condition.*

Condition	Psychological flexibility *M* (*SD*)	Generalised pliance *M* (*SD*)	Cognitive fusion *M* (*SD*)	Analogical reasoning *M* (*SD*)	Age *M* (*SD*)	Sex (female); %
A	22.81 (7.53) [*n*=21]	66.29 (18.43) [*n*=21]	25.38 (8.06) [*n*=21]	22.71 (4.36) [*n*=21]	23.43 (5.11) [*n*=21]	*n*=17; 81%
B	22.25 (9.48) [*n*=20]	63.30 (19.05) [*n*=20]	23.90 (12.01) [*n*=20]	19.76 (3.90) [*n*=21]	24.30 (7.87) [*n*=20]	*n*=17; 81%
C	22.86 (6.92) [*n*=21]	65.57 (23.00) [*n*=21]	24.95 (7.50) [*n*=21]	22.61(2.94) [*n*=21]	25.86 (10.81) [*n*=21]	*n*=16; 76.2%
D	23.76 (8.64) [*n*=21]	66.43 (19.79) [*n*=21]	26.67 (10.29) [*n*=21]	22.71 (4.85) [*n*=21]	26.57 (11.09) [*n*=22]	*n*=16; 76.2%

Note. A = Common properties and augmentals, B = Common properties only; C = Augmentals only; D = Neither

Table 2. *Means and SDs for pain tolerance and pain strength scores per time per condition, while controlling for analogical reasoning*.

Pain tolerance					Pain strength				
Condition	Time 1		Time 2		Condition	Time 1		Time 2	
	M	(*SD*)	*M*	(*SD*)		*M*	(*SD*)	*M*	(*SD*)
A (*n*=21)	31.53	21.02	49.3	59.97	A (*n*=14)	6.36	1.74	6.00	2.03
B (*n*=21)	67.19	74.72	93.42	90.77	B (*n*=13)	6.85	2.46	5.89	2.26
C (*n*=21)	34.18	16.77	73.36	77.69	C (*n*=19)	6.82	1.75	6.68	2.11
D (*n*=21)	50.61	41.27	83.83	75.89	D (*n*=16)	6.39	2.44	5.66	2.53

Note. A = Common properties and augmentals, B = Common properties only; C = Augmentals only; D = Neither

Table 3. *Moderations in PROCESS (bootstrapped samples 5000).*

Moderator		B	SE B	*p*	95% CI for B Lower	Upper
Psychological inflexibility	Constant	33.57	11.89	0.01	9.87	57.26
	Common Properties Only	-8.03	17.01	0.64	-41.91	25.85
	Augmentals Only	5.53	16.78	0.74	-27.89	38.95
	Both	-15.62	16.78	0.35	-49.05	17.80
	Psychological Inflexibility	-0.41	1.40	0.77	-3.21	1.38
	Common Properties * Psych Inflex	-2.03	1.92	0.30	-5.86	1.80
	Augmentals * Psych Inflex	-0.69	2.24	0.76	-5.16	3.78
	Both * Psych Inflex	1.93	2.14	0.37	-2.33	6.19
Cognitive fusion	Constant	32.49	12.06	0.01	8.47	56.52
	Common Properties Only	-7.02	17.23	0.68	-41.35	27.31
	Augmentals Only	7.21	16.97	0.67	-26.61	41.02
	Both	-14.88	16.97	0.38	-48.69	18.92
	Cognitive Fusion	0.51	1.12	0.67	-1.86	2.88
	Common Properties * Cognitive Fusion	-1.79	1.58	0.26	-4.94	1.37
	Augmentals * Cognitive Fusion	1.32	2.02	0.52	-2.70	5.34
	Both * Cognitive Fusion	0.57	1.93	0.77	-3.27	4.41

Table 4. *Moderations in PROCESS (bootstrapped samples 5000).*

Moderator		B	SE B	*p*	95% CI for B Lower	Upper
Generalised Pliance	Constant	32.68	12.08	0.01	8.62	56.74
	Common Properties Only	-6.98	17.34	0.69	-41.52	27.56
	Augmentals Only	6.49	17.07	0.70	-27.51	40.49
	Both	-14.91	17.08	0.39	-48.94	19.11
	Generalised Pliance	0.54	0.62	0.39	-0.71	1.78
	Common Properties * Generalised Pliance	-1.24	0.91	0.18	-3.05	0.58
	Augmentals * Generalised Pliance	-0.52	0.82	0.53	-2.16	1.12
	Both * Generalised Pliance	-0.54	0.92	0.56	-2.37	1.29
Analogical Reasoning	Constant	32.45	11.59	0.01	9.36	55.54
	Common properties	-2.37	17.58	0.89	-37.38	32.63
	Augmentals only	-0.48	16.51	0.98	-33.36	32.4
	Both	-17.28	16.42	0.3	-49.98	15.43
	Analogical Reasoning	1.01	2.42	0.68	-3.81	5.83
	Common Properties * Analogical Reasoning	0.75	3.86	0.86	-6.95	8.43
	Augmentals * Analogical Reasoning	9.79	4.66	0.04	0.5	19.08
	Both * Analogical Reasoning	2.39	3.62	0.51	-4.82	9.59